Available online at www.sciencedirect.com
SCIENCE DIRECT
Analytica Chimica Acta 553 (2005) 134–140
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

# The use of the autocorrelation function in modeling of multivariate data

C.E. Alciaturi *, M.E. Escobar, I. Estéves

*Instituto Zuliano de Investigaciones Tecnológicas (INZIT-CICASI), Apartado Postal 331, Maracaibo, Venezuela*

Received 27 February 2005; received in revised form 16 July 2005; accepted 1 August 2005
Available online 12 September 2005

**Abstract**

The use of the autocorrelation function ($R_1$) with lag 1 in univariate and multivariate model selection is proposed. In the univariate case, a $Z$-value is calculated from the $R_1$ of the residual vector. A high positive value of $Z$ indicates the presence of smooth, non-random variations in the data not explained by the model considered. In the multivariate case, a new procedure is proposed. A "short" path is found in the independent variable space, and the $Z$ from the dependent variable residual vector is used to measure the smoothness of the changes in the dependent variable. Applications are shown for variable selection and determination of the number of latent variables in PLS1 models.


*Keywords:* Autocorrelation; Smoothness; Model selection; Linear models

## 1. Introduction

The application the autocorrelation function of residuals at lag 1 ($R_1$) and the resulting $Z$-values to curve fitting was shown in a previous paper [1]. It was demonstrated that, that when fitting data with random noise, $R_1$ provided a way of rejecting approximations, and of making comparisons between various models. Large positive or negative values of $R_1$ caused the rejection of models, while small values of $R_1$ are consistent with the presence of random noise.

In this paper, we propose to use the $R_1$ of the residuals of linear models as an indication of the presence (or not) of "smoothness" in the residual. If the residual of a linear model is smooth, it may be possible to model it by a non-linear function, or to select a different linear model.

It is often assumed that physical or chemical properties in a neighborhood of space or in an interval of time present some coherence and generally do not change abruptly. Although there are exceptions (chaotic systems and phase transitions), this is generally the case with the systems in analytical or process chemistry. When trying to predict the values of a dependent variable from an independent variable, it is highly desirable to have a "smooth" relationship between both of them. Conversely, if there is no "smoothness", the presence of random noise is highly likely. "Smoothness" has been defined in various forms. For the univariate case, the smoothness potential of a function may be defined by:

$$U(f) = \int |f^{(n)}(x)|^2 \, \mathrm{d}x \tag{1}$$

where $n$ is the order of the derivative. When the purpose is to assess the smoothness of data, the derivatives may be estimated by numerical methods. Thus, for a pair of vectors $\mathbf{x}$ and $\mathbf{y}$, where $\mathbf{y}$ represents the dependent variable and $\mathbf{x}$ is equally spaced, the smoothness (or, rather the roughness) may be estimated by the following expression (see, for example, [2]):

$$\text{Roughness} = \sum \left| \frac{\mathbf{y}_{j+1} - 2\mathbf{y}_j + \mathbf{y}_{j-1}}{h} \right|^2 \tag{2}$$

where $h$ is the increment in $\mathbf{x}$. (The terms in the summation estimate the second-derivative.) Some applications of smoothness in regression have been discussed in Ref. [2]. Eq. (2) may be readily applied to equally spaced data, as is often the case with time series and infrared spectra. If the spacing of $\mathbf{x}$ were irregular (as when $\mathbf{x}$ represents a physical

* Corresponding author. Fax: +58 61 913769.
*E-mail address:* alciaturi@hotmail.com (C.E. Alciaturi).

0003-2670/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2005.08.001

property as absorbance or pH), a numerical estimation of derivatives with Eq. (2) would give unstable results.

Programs were written in-house using Matlab 6.5 in a personal computer with a Pentium 4 processor.

## 2. The autocorrelation function for estimation of smoothness

### 2.1. One independent variable

A function that should be more convenient for the analysis of irregularly spaced data is the autocorrelation function $R_L$ [1] defined as the linear correlation of a series of values with the same series with a specified lag. This function has limits +1 and −1. Thus, a smooth series will have $R_L$ values close to +1, a random series close to 0, and an alternating series close to −1. The expected standard deviation in a series is given by:

$$\sigma_L = \left(\frac{1}{n-L}\right)^{1/2} \tag{3}$$

where $n$ is the series length and $L$ the lag. Also, $Z$-values may be calculated, and a probability associated with each observation [3]:

$$Z = \frac{R_L - 0}{\sigma_L} \tag{4}$$

Here, we propose to use the autocorrelation function with lag 1 ($R_1$) for detection of smooth relationships. A high positive value of $R_1$ (or $Z$) will indicate the presence of smooth variations in the **y** vector. If the vector is a residual, it may be further modeled, for example, by a non-linear function.

When there are many independent variables (as is FTIR or NIR spectroscopy) this procedure may be helpful for variable selection, as will be shown in the examples. An algorithm is proposed for multiple independent variables in the next section.

### 2.2. The algorithm for two or more independent variables

The following procedure is proposed for making extensions to two or more independent variables (calculation of "multivariate" $R_1$ and $Z$):

(a) If necessary, normalize each of the independent variables.
(b) Find a sequence of points (in the space of the independent variables) such that the distances between consecutive points are small. Our algorithm starts with the closest pair of points and then finds the next closest point and so on, constructing a "line" of points until all are connected. This sequence gives **yc**, a vector of values for the dependent variable, and **xc**, a vector of cumulative distances. The values of the elements of **yc** are the same of **y**, but with a different order, determined by the order in **xc**.
(c) Plot an **xc**–**yc** graph of the values of the dependent variable versus cumulative distance of the independent variable.
(d) Calculate the autocorrelation with lag 1 and the $Z$ of the **yc** values.

The **xc**–**yc** graph will show if the relation is "smooth", i.e. whether a small variation in **xc** is corresponded with a small change in **yc**. The presence of a smooth relation will also be shown by a high positive value of $Z$.

### 2.3. A procedure for model selection with projection to latent structures (PLS1) or principal components regression (PCR)

For a given **X** (matrix of independent variables, where the columns are variables and the rows samples) and **y** (vector of the dependent variable, representing a property of the samples):

(a) Apply the algorithm of Section 2.2. This will give a sequence of points (samples), an **xc** vector in the independent variable space, a **yc** vector of property values, and a $Z$-value. A high positive value of $Z$ will indicate a "smooth" relationship between **X** and **y**, so it should be possible to predict **y** from **X**.
(b) Do PLS1 (or PCR) for a selected number of latent variables (or principal components), and find $\mathbf{y}_{res}$, the residual of **y**.
(c) Repeat part (a), with the same **X** matrix, sequence of points, and **xc** vector, but with $\mathbf{y}_{res}$ from part (b), which gives a $\mathbf{yc}_{res}$. Calculate the value of $Z$(residual).

A value of $Z$(residual) close to zero will indicate that **X** contains no information related to the residual of **y**. Thus, this procedure may be applied for selecting the "right" number of latent variables (or principal components).

It should be pointed out that a similar test (the Durbin–Watson statistic, which also considers successive points) has been used by Rutledge and Barros [4] to determine the regression model dimensionality for several examples in PLS1 regression. The DW statistic gives values close to zero when there is a strong positive correlation between successive points, and close to two for a random distribution. Their procedure involved applying the DW criterion to the ***p***, ***w***, and ***b*** vectors to detect latent variables with low signal/noise ratios.

## 3. Results and discussion

### 3.1. Synthetic univariate data

Fig. 1 shows two cases: a logarithmic function (a) and a first-order polynomial with random noise added (b), both with